Math 455 Class 37 April 30

Midterm 2 will be graded by Monday (hopefully!)

Remarks on a problem from the practice midterm on Q-Q plots.

From Faraway p. 84, on "standardized residuals".

"If the model assumptions are correct, $\text{var}(r_i) = 1$ and $\text{cor}(r_i, r_j)$ tends to be small."

The ideal Q-Q plot shows the dots on the line $y = x$, not indicating any failure of the model assumptions.

Why is this so? An experimental investigation with code.

Suppose we generate 50 standard normal RVs (iid), sort them, and do a Q-Q plot.

We expect the dots to be on the line $y = x$.

Because the i^{th} sorted number ~~the~~ has the distribution of the i^{th} order statistic. See Wackerly Ch. 6 for theoretical distribution of the i^{th} order stat.

The i^{th} order statistic has a very sharply peaked distribution.

~~Intuitio~~ Intuition: the k^{th} order statistic should be at (approximately) the $\frac{k}{n}$ percentile of the distribution.

(There are more detailed results possible.)

This requires that our samples be independent.

If the samples are not independent, then we will NOT get the expected result in our Q-Q plot.

We looked at R code showing examples of this.

Easy to generate such code with LLM.

WARNING: If you are interviewing for coding work, make sure you understand the code your LLM is generating.

Problem from Monday's midterm.

(#3) the solution is "need I(X^2)".

This is a typical "gotcha" or "explain what's going on" question from coding interviews.

Other examples: do this LeetCode problem right now in a live interview and explain your thinking.

- Explain all the uses of "const" in this line of C++ code.

Explaining the picture on the front cover of the textbook: Ch. 11. "Shrinkage Methods".

One symptom of collinearity: large standard errors for coefficient estimates. Large and unstable values of coefficient estimates.
(Examples of this in Ch. 7.)

Ch 7. addresses underlying collinearity.

One way of dealing with this is to remove some of the variables, or replace variables with an "index" created from several variables.

But we might have so many variables we can no longer make such selections by hand.

Another way of dealing with this: treat the symptom directly:

If the problem is "large and unstable coefficient estimates," let's penalize large coefficients.

Instead of minimizing $SSR = \sum_{i=1}^{n} e_i^2$

minimize: SSR + penalty term.

Two typical penalty terms:

"Ridge" and "LASSO"

Ridge: $\lambda\left(\hat{\beta}_1^2 + \cdots + \hat{\beta}_p^2\right)$

LASSO: $\lambda\left(|\hat{\beta}_1| + \cdots + |\hat{\beta}_p|\right)$.

Notice $\hat{\beta}_0$ is NOT (typically) included in either penalty term.

These penalty terms explain the picture on the front cover.

ICBST: penalty terms are equivalent to "constrained regression"

Another way of making the coefficient estimates smaller is to say, "let's minimize SSR subject to the constraint

$$\hat{\beta}_1^2 + \cdots + \hat{\beta}_p^2 \leq K \qquad \text{(Ridge)}$$

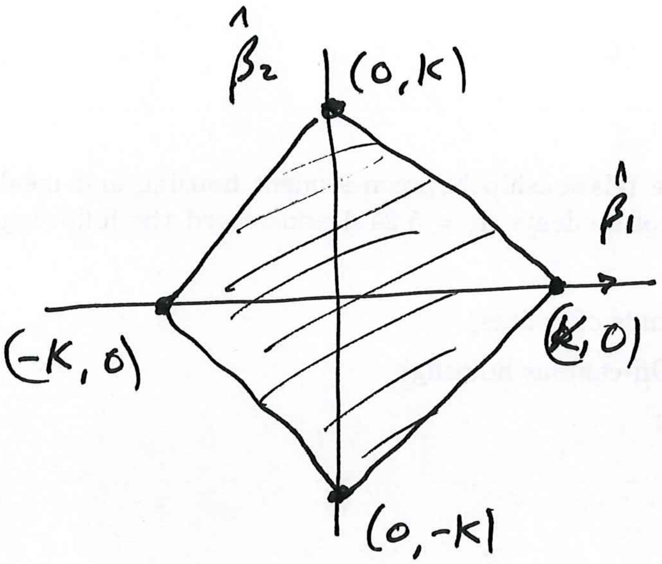
or $$|\hat{\beta}_1| + \cdots + |\hat{\beta}_p| \leq K. \qquad \text{(Lasso)}$$

this is equivalent to the penalty term in the sense that for any choice of λ, there is a corresponding choice of K that gives the same result.

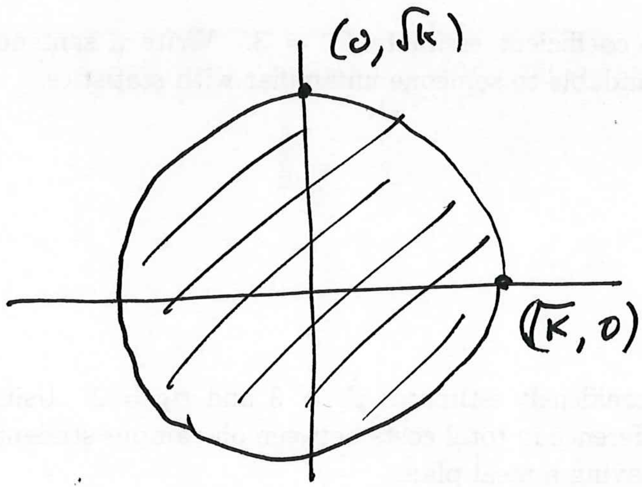
Explaining the picture: let $p=2$, so we can draw something in the $\hat{\beta}_1, \hat{\beta}_2$ plane.

Let's draw the regions:

$$|\hat{\beta}_1| + |\hat{\beta}_2| \leq K \qquad \hat{\beta}_1^2 + \hat{\beta}_2^2 \leq K.$$



Shaded region is $|\hat{\beta}_1| + |\hat{\beta}_2| \le K$.



Shaded region is $\hat{\beta}_1^2 + \hat{\beta}_2^2 \le K$.

What point ~~will be~~ $(\hat{\beta}_1, \hat{\beta}_2)$ will be selected by the constrained minimization procedure?

Picture on front cover shows unconstrained minimization & confidence ellipses with increasing SSR (constant on each ellipse).